中国计算机学会学术著作丛书

知识发现（第二版）

Knowledge Discovery (Second Edition)

史忠植 著

清华大学出版社

中国计算机学会学术著作丛书

知识发现（第二版）

Knowledge Discovery
(Second Edition)

史忠植　著

清华大学出版社
北京

内 容 简 介

知识发现是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。知识发现将信息变为知识，从数据资源中发现知识宝藏，将为知识创新和知识经济的发展作出贡献。

本书全面而又系统地介绍了知识发现的方法和技术，反映了当前知识发现研究的最新成果和进展。全书共分 15 章。第 1 章是绪论，概述知识发现的重要概念和发展过程。下面三章重点讨论分类问题，包括决策树、支持向量机和迁移学习。第 5 章阐述聚类分析。第 6 章是关联规则。第 7 章讨论粗糙集和粒度计算。第 8 章介绍神经网络，书中着重介绍几种实用的算法。第 9 章探讨贝叶斯网络。第 10 章讨论隐马尔可夫模型。第 11 章探讨图挖掘。第 12 章讨论进化计算和遗传算法。第 13 章探讨分布式知识发现，它使海量数据挖掘成为可能。最后两章以 Web 知识发现、认知神经科学为例，介绍知识发现的应用。

本书内容新颖，认真总结了作者的科研成果，取材国内外最新资料，反映了当前该领域的研究水平。论述力求概念清晰，表达准确，算法丰富，突出理论联系实际，富有启发性。

本书可以用作高等院校有关专业的研究生和高年级本科生的知识发现、数据挖掘、机器学习等课程教材，也可供从事知识发现、数据挖掘、机器学习、智能信息处理、模式识别、智能控制研究和知识管理的科技人员阅读参考。

图书在版编目(CIP)数据

知识发现/史忠植著. --2 版. --北京：清华大学出版社，2011.1

(中国计算机学会学术著作丛书)

ISBN 978-7-302-23957-4

Ⅰ. ①知… Ⅱ. ①史… Ⅲ. ①知识工程 ②人工智能 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字(2010)第 199169 号

责任编辑：薛　慧

责任校对：刘玉霞

责任印制：王秀菊

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座

http://www.tup.com.cn　　**邮　　编**：100084

社　总　机：010-62770175　　**邮　　购**：010-62786544

投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn

质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 装 者：清华大学印刷厂

经　　销：全国新华书店

开　　本：175×245　**印　张**：32.25　**字　数**：625 千字

版　　次：2011 年 1 月第 2 版　　**印　　次**：2011 年 1 月第1 次印刷

印　　数：1～3000

定　　价：59.00 元

产品编号：035243-01

序

Preface

第一台电子计算机诞生于20世纪40年代。到目前为止，计算机的发展已远远超出了其创始者的想象。计算机的处理能力越来越强，应用面越来越广，应用领域也从单纯的科学计算渗透到社会生活的方方面面：从工业、国防、医疗、教育、娱乐直至人们的日常生活，计算机的影响可谓无处不在。

计算机之所以能取得上述地位并成为全球最具活力的产业，原因在于其高速的计算能力、庞大的存储能力以及友好、灵活的用户界面。而这些新技术及其应用有赖于研究人员多年不懈的努力。学术研究是应用研究的基础，也是技术发展的动力。

自1992年起，清华大学出版社与广西科学技术出版社为促进我国计算机科学技术与产业的发展，推动计算机科技著作的出版，设立了“计算机学术著作出版基金”，并将资助出版的著作列为中国计算机学会的学术著作丛书。时至今日，本套丛书已出版学术专著近50种，产生了很好的社会影响，有的专著具有很高的学术水平，有的则奠定了一类学术研究的基础。中国计算机学会一直将学术著作的出版作为学会的一项主要工作。本届理事会将秉承这一传统，继续大力支持本套丛书的

出版,鼓励科技工作者写出更多的优秀学术著作,多出好书,多出精品,为提高我国的知识创新和技术创新能力,促进计算机科学技术的发展和进步作出更大的贡献。

中国计算机学会

2002 年 6 月 14 日

前言

Foreword

知识发现将从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。它集数据收集、数据清洗、降维、规则归纳、模式识别、结果分析与评估、可视化输出等多种功能于一身，是统计学、机器学习、神经网络、计算机科学、模式识别、人工智能及其他学科相结合的产物。

从20世纪80年代中期的小范围研究到如今的蓬勃发展，知识发现已经在企业界与科学界占据了一席之地。在信息管理、查询响应、决策支持、过程控制、搜索引擎、基因分析等许多方面得到了广泛应用。帮助企业进行客户关系管理，优化生产过程，减少不必要的投资，提高资金周转和回报。帮助人们迅速获取所需的知识和信息，提高工作效率，改进服务质量。不同领域的人们都运用知识发现，将信息变为知识，从数据资源中发现知识宝藏。

本书第一版自2002年出版以来，国内外知识发现的研究者们取得了许多重要进展，涌现出了许多新概念、新方法、新算法，应用领域也日新月异。本书第二版对原书作了重大修改，增加迁移学习、聚类分析、图挖掘、分布式知识发现、认知神经科学知识发现等内容，尽可能反映这近十年间知识发现、数据挖掘，及其应用所取得的新成果，指

出当今活跃的研究方向和课题,给读者有所启发。

本书全面而又系统地介绍知识发现的方法和技术。全书共分 15 章。第 1 章是绪论,概述知识发现的重要概念和发展过程。第 2 章讨论决策树,它是归纳学习方法中最实用的一种技术,新版中增加决策树并行挖掘算法。广泛应用的支持向量机在第 3 章讨论,增加了改进算法。第 4 章探讨迁移学习。前面三章重点讨论分类问题。第 5 章阐述聚类分析。第 6 章是关联规则,它是近几年应用最为广泛的挖掘方法之一。第 7 章讨论粗糙集,增加了粒度计算的内容。第 8 章介绍神经网络,书中着重介绍几种实用的算法。第 9 章探讨贝叶斯网络,贝叶斯网络可以处理不完整和带有噪声的数据集,它用概率测度的权重来描述数据间的相关性。它在近几年知识发现研究中是极其活跃的研究课题。第 10 章讨论隐马尔可夫模型。第 11 章探讨图挖掘,这是近期极其活跃的研究领域,在互联网分析、社会计算、生物信息学等方面得到广泛应用。第 12 章讨论进化计算和遗传算法。第 13 章探讨分布式知识发现,它使海量数据挖掘成为可行,重点阐述多主体(multi-agent)、网格计算、云计算环境下的海量数据挖掘。最后以 Web 知识发现、认知神经科学为例,介绍知识发现的应用。第 14 章关于 Web 知识发现。第 15 章介绍认知神经科学中如何运用知识发现技术分析认知神经科学的实验结果,探索智能的本质和机理。

本书总结了作者和中国科学院计算技术研究所智能科学实验室多年的科研成果,也吸取了国内外同行的研究成果和有关文献的精华,在此谨向这些成果和文献的作者表示感谢,他们的丰硕成果和贡献是本书学术思想的重要源泉。本书的顺利撰写离不开中国科学院计算技术研究所智能科学实验室同事们的贡献和支持,特别是何清、叶世伟、李晓黎、叶施仁、宫秀军、刘少辉、郑毅、郑金华、胡宏、张建、王军、张颖、潘谦红、吴斌、贾自艳、秦亮曦、游湘涛、任力安、李宝东、梁吉业、丁世飞、蒙祖强、施智平、李清勇、郑征、罗平、罗杰文、石志伟、石川、李志清、李志欣、马慧芳、庄福振、刘曦、崔志华、周菁等给予了帮助和支持,在此一并表示感谢。

本书研究工作得到自然科学基金重点项目“基于云计算的海量数据挖掘”(批准号:61035003)、“基于感知学习和语言认知的智能计算模型研究”(批准号:60435010)、“Web 搜索与挖掘的新理论与方法”(批准号:60933004)、自然科学基金项目“语义 Web 服务的逻辑基础”(批准号:60775035)等的支持。感谢国家重点基础研究发展计划“基于视觉认知的非结构化信息处理理论与关键技术”(项目编号:2007CB311000)、国家 863 高技术探索项目“软件自治愈与自恢复技术”(项目编号:2007AA01Z132)等项目的支持。

本书内容新颖,反映了当前该领域国内外的研究水平。论述力求概念清晰,表达准确,算法丰富,突出理论联系实际,通过实例说明原理,富有启发性。

本书可以用作高等院校有关专业的研究生和高年级本科生的知识发现、数据挖掘、机器学习等课程教材，也可供从事知识发现、数据挖掘、机器学习、智能信息处理、模式识别、智能控制研究和知识管理的科技人员阅读参考。

由于作者水平有限，加上知识发现发展很快，研究领域广泛，书中不妥和错误之处在所难免，恳请各位专家和广大读者不吝指教和帮助。

知识就是力量，知识是创新的源泉。

史忠植

2011 年 1 月

目 录

Contents